What are Probability Survey Designs?

Steve Paulsen
Tony Olsen
Office of Research and Development
U.S. Environmental Protection Agency

Types of Statistical Designs

- **Experimental designs**
  - Random allocation of treatments
- **Observational studies**
  - Factor space designs
    - Gradient studies
  - Available sites
- **Survey designs**
  - Census
  - Probability survey
- **Response designs needed in all**

Survey Design : Response Design

- **Survey design is process of selecting sites at which a response will be determined**
  - Probability model for inference is based on the randomized selection process
  - Has a spatial component and may have a time component
- **Response design is process of obtaining a response at a site:**
  - A single index period during a year
  - Multiple periods during year: monthly, quarterly

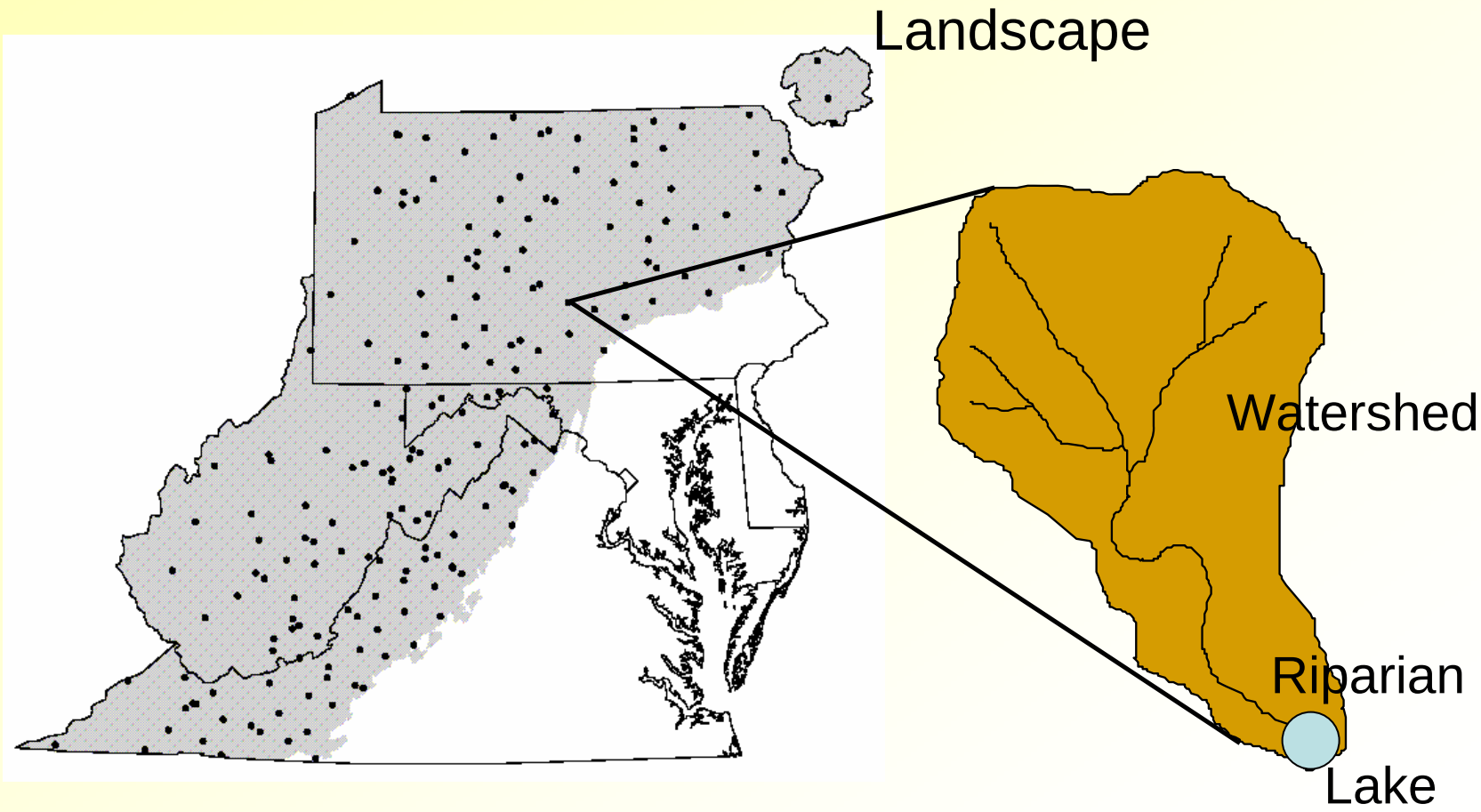
Response Design - What is It?

- Once you have selected a site to visit, how do you sample it for the selected indicators?
- Response design can have both a temporal and a spatial dimension.
- Requires defining the “target population” for which the design is applicable.
- Ultimately, it includes how you collapse the measurements into and “indicator”

The Response Design:
Index Period

- Time period within year selected for measurement (ecologically based)
- Measurements may be taken more than once during index period with response design giving protocol for obtaining single value for indicator
- Indicator variability within index period contributes to non-survey sampling error

Lake Study Design: Response Design Scales



Basic Spatial Survey Designs

- **Simple Random Sample**
- **Systematic Sample**
  - Regular grid
  - Regular spacing on linear resource
- **Spatially Balanced Sample**
  - Combination of simple random and systematic
  - Guarantees all possible samples are distributed across the resource (target population)
  - Generalized Random Tessellation Stratified (GRTS) design

Why Aren't Basic Designs Sufficient?

- Monitoring objectives may include requirements that basic designs can't address efficiently
  - Estimates for particular subpopulations requires greater sampling effort
  - Administrative restrictions and operational costs
- Ecological resource occurrence in study region makes basic designs inefficient
  - Resource is known to be restricted to particular habitats

Sample Survey Design

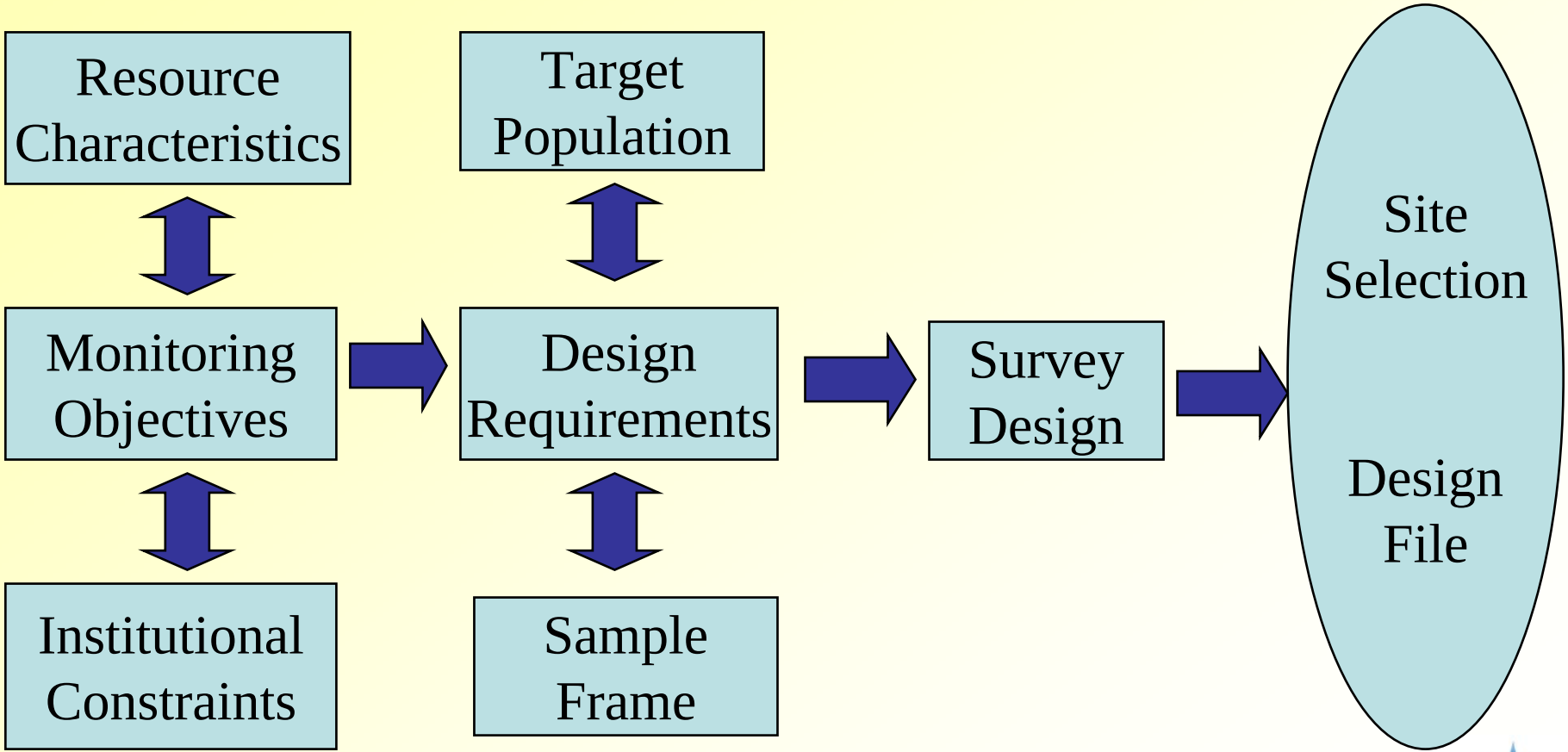
Probability samples for ecological resources with:

- spatial balance
- variable inclusion probabilities
- stratification
- variable density (for special interest areas)
- nested sub-sampling
- panel designs (for surveys over time)
- over-sampling (to account for frame errors)
- confidence intervals

For multiple resource types:

- discrete resources (lakes; wetlands)
- areal resources (estuaries; large lakes)
- continuous resources (streams and rivers)

Design Structure



Ecological Resource Characteristics From Survey Design Perspective

- **Finite population of discrete entities**
  - 0-dimensional
  - All small lakes in the 48 conterminous states
  - All 8-digit USGS CU in the 48 conterminous states
- **Continuous areal population**
  - 2-dimensional
  - All forest land
  - All coastal estuarine resources
- **Continuous linear network population**
  - 1-dimensional embedded in 2-dimensions
  - All perennial wadeable streams

Institutional Constraints

- Budget
- Field crews: location and number
- Separate biological and water quality groups
- Integration of 305(b) and 303(d) monitoring

Monitoring Objectives

- Objectives determine the monitoring design
  - Usual to have multiple objectives
  - Objectives compete for samples
  - Precise statements are required
  - Objectives must be prioritized
- Target population and subpopulations are determined by objectives

What is a Target Population?

- Target population denotes the ecological resource about which information is wanted
- Requires a clear, precise definition
  - Must be understandable to users
  - Field crews must be able to determine if a particular site is included
- More difficult to define than most expect.
- Includes definition of what the elements are that make up the target population

Subpopulations and Domains

- Subsets of the target population that are of particular interest
- Examples for aquatic ecosystems
  - Ecoregions, biogeographic regions
  - All lentic resources in region with area < 100 ha
  - All lotic resources with with Strahler order < 4
  - Tidal creeks versus open water estuarine areas
  - All lotic resources with < 20% riparian canopy cover
  - All 5-th field HUCs with >10 NWI wetland polygons
  - All 6-th field HUCs with >25% Federal land ownership

Subpopulations: Impact on Design

- Objectives identify critical subpopulations with expected sample sizes: Domains
- Survey design addresses domain sample size requirements
  - Explicitly using stratification, unequal weighting
  - Implicitly when other requirements provide sufficient sample sizes
- Other subpopulations can not be defined prior to sample selection

Sample Frame

- What is it and why is it important
- GIS representation of the target population
- Issues:
  - Overcoverage: includes elements not in target
  - Undercoverage: excludes elements not in target
  - Coding errors
- NHD for streams, lakes, estuaries

Design Requirements

- Target population definition
- Subpopulations for which estimates are required
- Unequal weighting categories (eg stream order)
- Sample frame selection
- Sample size
- Single year or multiple year study

Survey Design Specification

- Target population definition
- Sample Frame Specification
- Basic Design
  - RTS for areal population
  - GRTS for linear continuous population
  - GRTS for discrete population
  - Simple random sample
- Stratification categories
- Multi-density categories
- Expected sample sizes
- Panel structure
- Oversample specification
- Nested Subsamples
- Intensification domains
- Study-wide base sample
- Multi-Stage or Multi-phase

Site Selection Process

- Driven by survey design specification
- Create GIS coverage for boundary of study area
- Create a hierarchical grid covering study area
  - Ensure <1 point per grid cell
  - Ensure grid cell small enough to be included in small regions of study area
- Create GIS coverage for sample frame
- Intersect frame coverage with grid coverage
- Determine multi-density multiplicative factors for unequal probability
- Select sites
- Create Design File

Design File Contents:
Has all sites selected by survey design

- Site Identification
  - Site ID
  - Latitude/Longitude
  - Site Name
  - Sample Frame ID
  - County
  - Map names
- Auxiliary Frame Information
  - State
  - Omernik Ecoregion
  - Other
- Survey Design
  - Stratum
  - Multi-Density Category
  - Panel
  - Oversample
  - Nesting Identification
  - Expected Sample Size
  - Initial Weight

Clean Water Act Questions

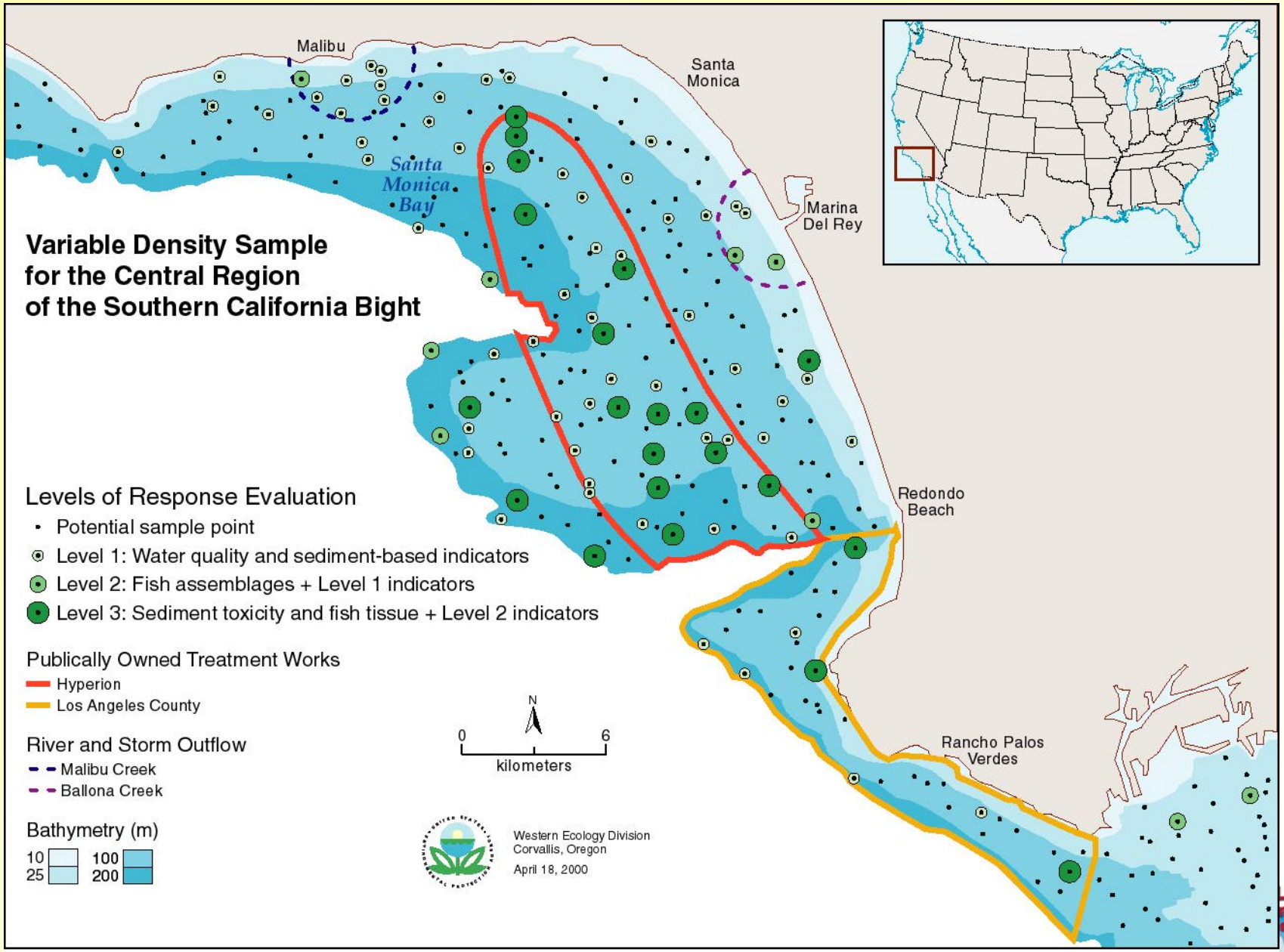
- Do I have a problem and if so how big???
  - What proportion of lakes exceed a particular TP threshold
- Is it getting better or worse???
  - Is the proportion of lakes exceeding TP threshold increasing or decreasing?
- What's causing the problem?
  - What's the relative importance of the stressors associated with the problem? (e.g., point versus non-point sources)

How do you get started?

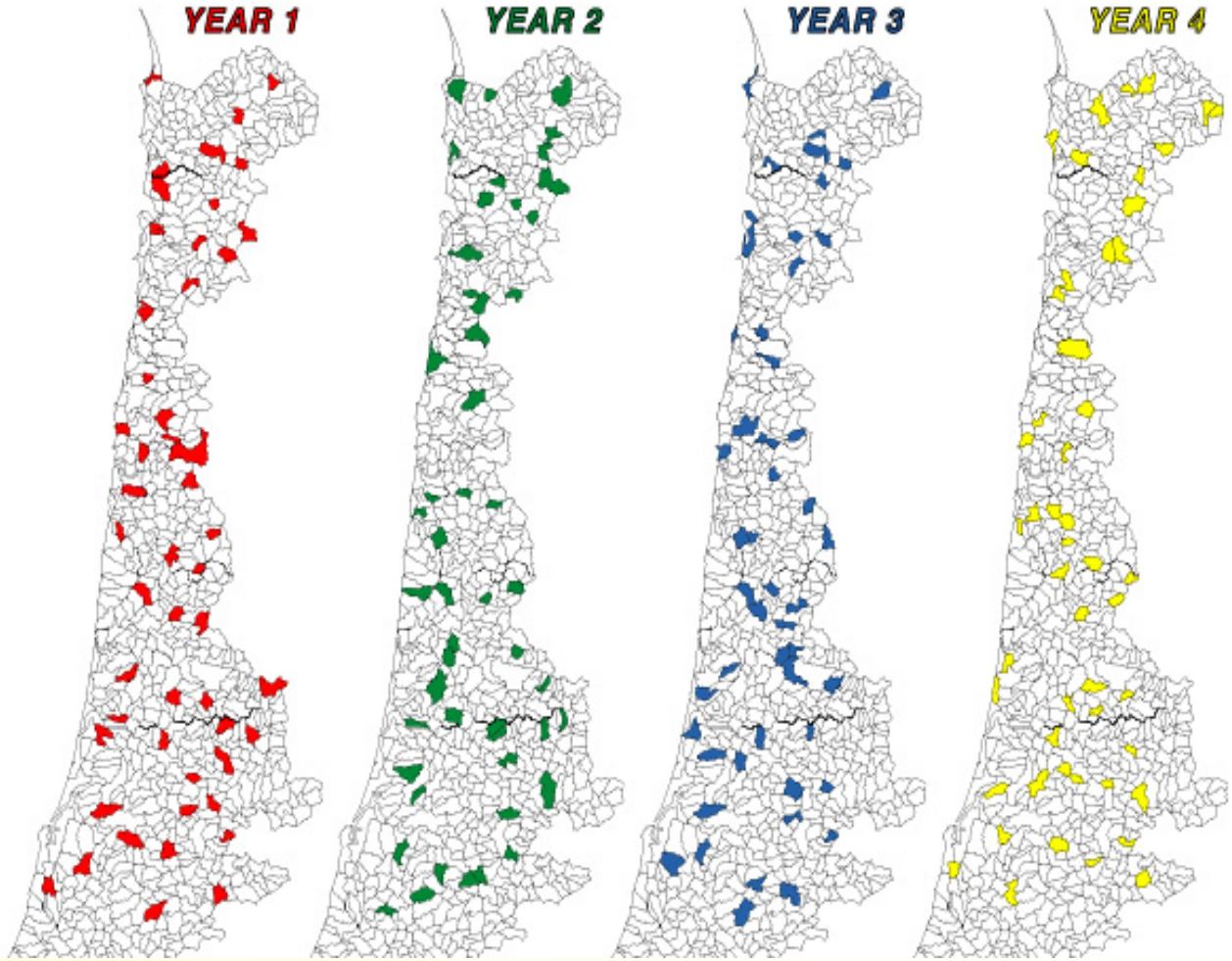
- What is your question? (duh)
- No, really what is your question? (*i.e.*, you need to be able to draw the graph you want to fill in and label the axes)
  - *e.g.*, What is the condition of the nation's lakes? This is a start but not sufficient.
  - How many lakes or what proportion of lakes in the lower 48 states have Secchi visibility below 2 meters?
  - How much of the lake acreage in the lower 48 states has Secchi visibility below 2 meters?
  - 2 different questions requiring 2 different designs.

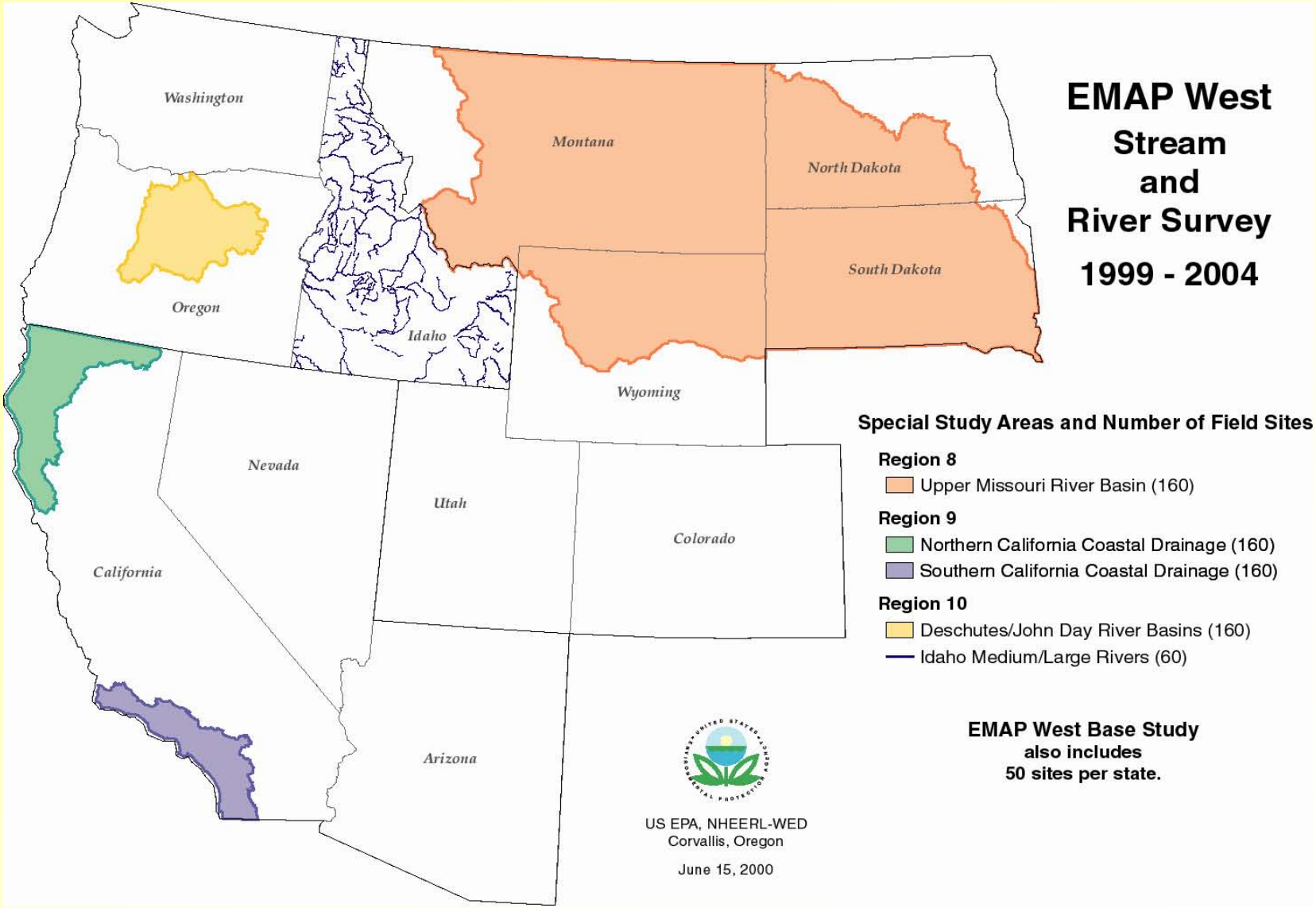
Additional Planning Questions

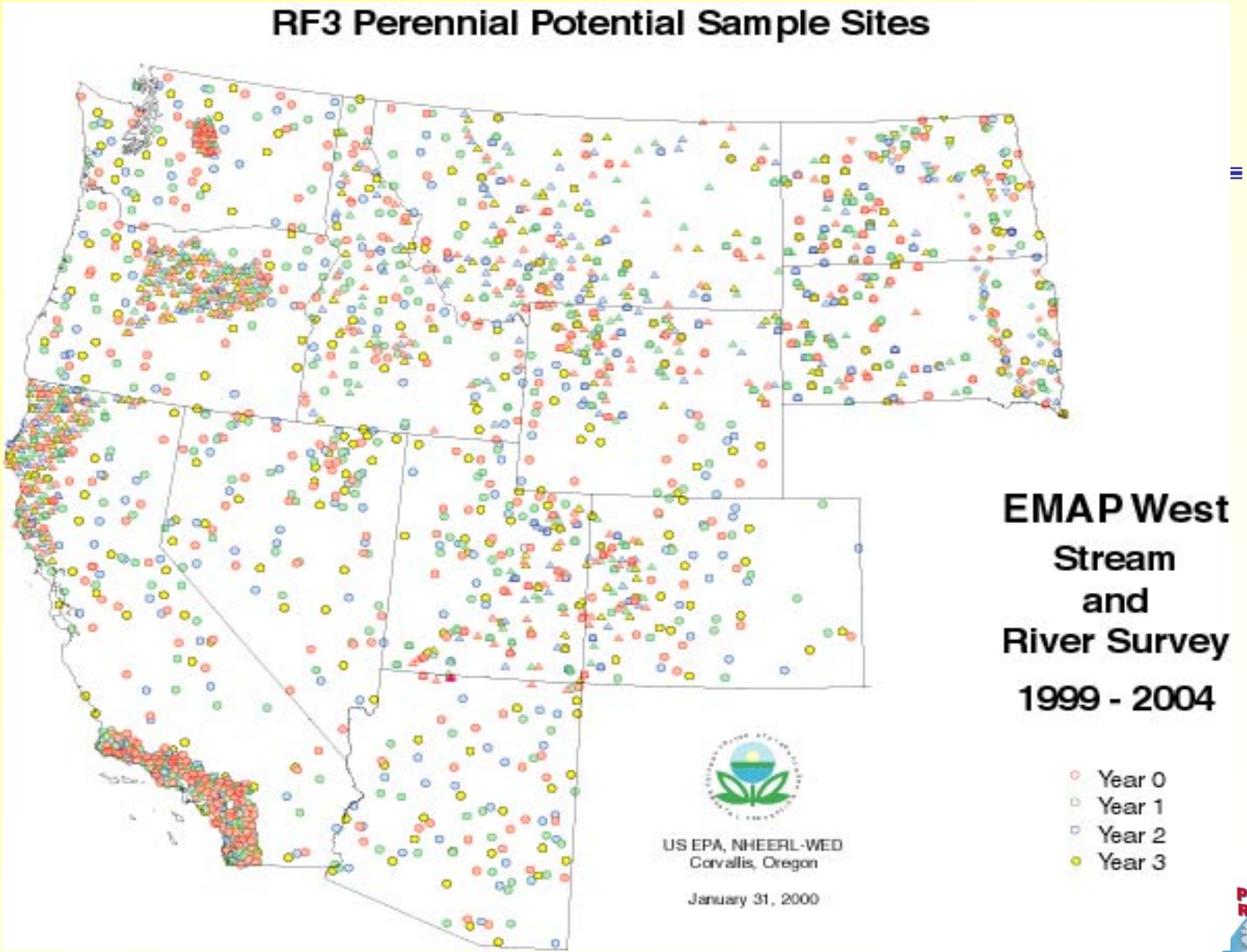
- How do you want to evaluate the results?
- Just one set of numbers for the entire population?
- Are there “sub-populations” that you want to compare?
  - Geographic subpopulations (*e.g.*, east – west)
  - Functional subpopulations (*e.g.*, size, stratified/non-stratified)
  - Only limited by creativity and sample size (aka $$$)
  - Can also “post classify” – can’t guarantee sample size


Malibu
Santa Monica
Santa Monica Bay
Marina Del Rey
Redondo Beach
Rancho Palos Verdes
Variable Density Sample for the Central Region of the Southern California Bight
Levels of Response Evaluation
Potential sample point
Level 1: Water quality and sediment-based indicators
Level 2: Fish assemblages + Level 1 indicators
Level 3: Sediment toxicity and fish tissue + Level 2 indicators
Publically Owned Treatment Works
Hyperion
Los Angeles County
River and Storm Outflow
Malibu Creek
Ballona Creek
Bathymetry (m)
10
25
100
200
N
0
6
kilometers
Western Ecology Division
Corvallis, Oregon
April 18, 2000
projects/emapgis/urquhart/california bight/figure12-4.ai
4/18/00 smp

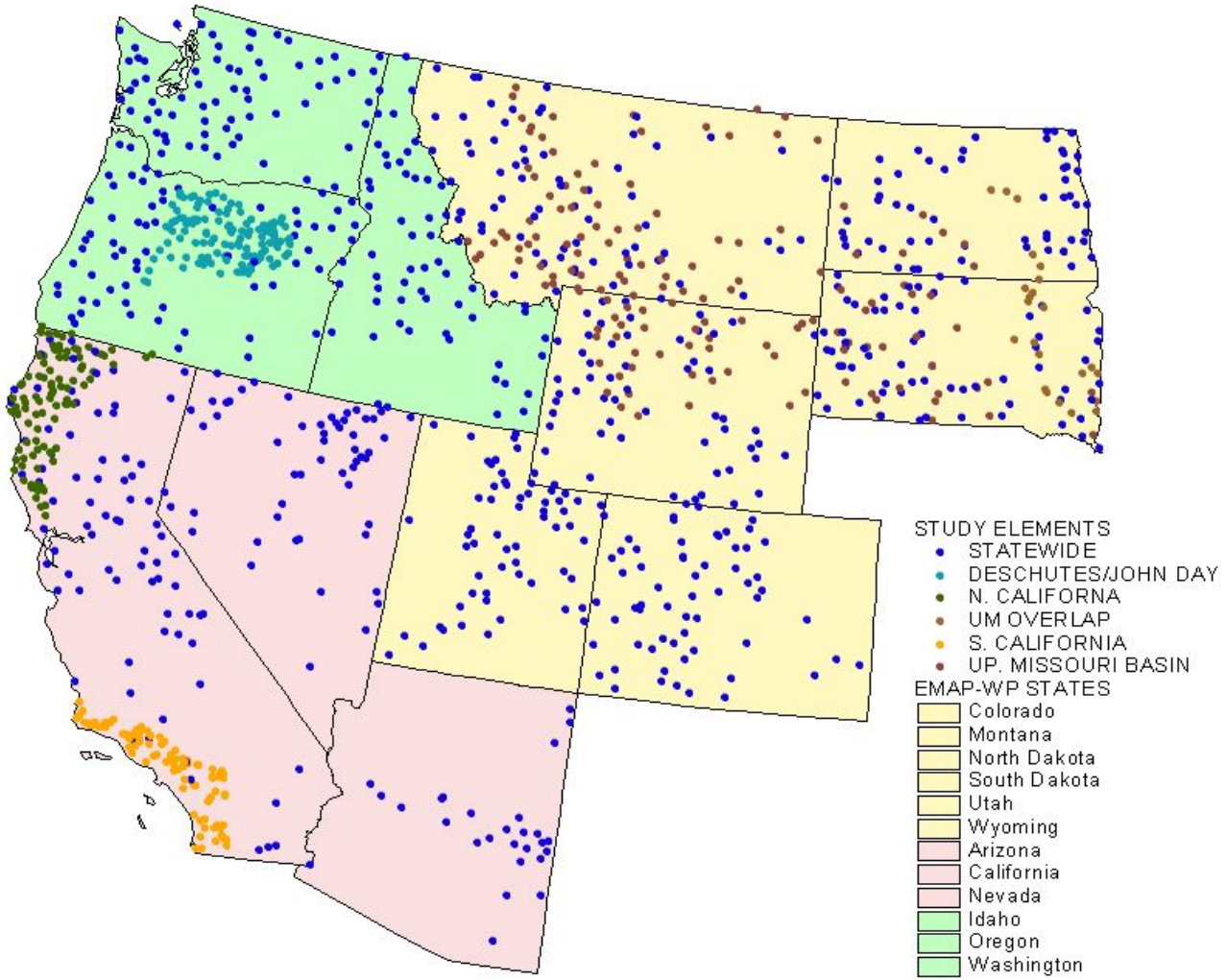
Spatially-Balanced Sample of 6-th Field Hydrologic Units
Coastal Region of Oregon



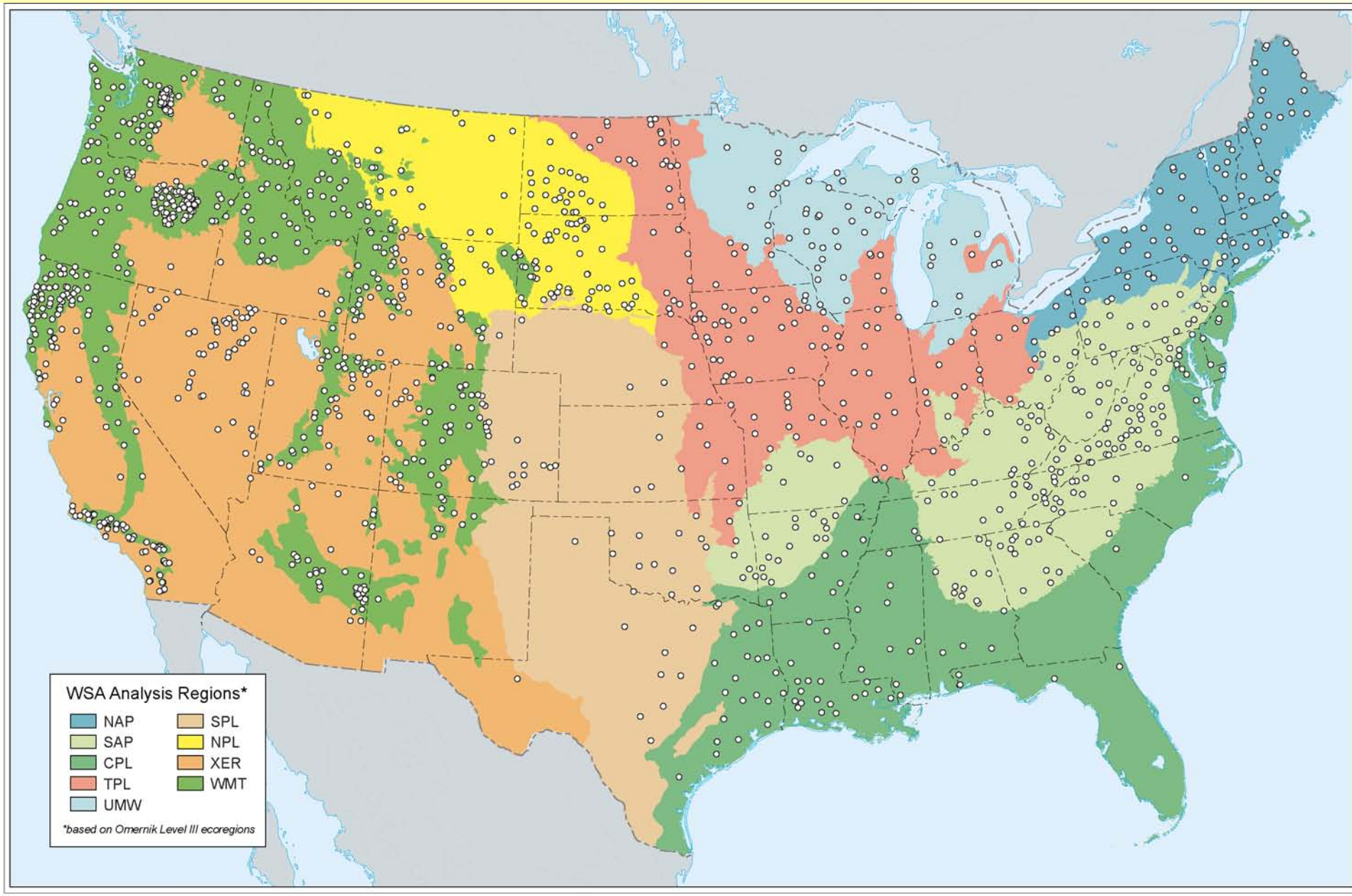

EMAP West
Stream
and
River Survey
1999 - 2004
Washington
Montana
North Dakota
South Dakota
Oregon
Idaho
Wyoming
Nevada
Utah
Colorado
California
Arizona
Special Study Areas and Number of Field Sites
Region 8
Upper Missouri River Basin (160)
Region 9
Northern California Coastal Drainage (160)
Southern California Coastal Drainage (160)
Region 10
Deschutes/John Day River Basins (160)
Idaho Medium/Large Rivers (60)
EMAP West Base Study
also includes
50 sites per state.
US EPA, NHEERL-WED
Corvallis, Oregon
June 15, 2000


RF3 Perennial Potential Sample Sites
EMAP West
Stream
and
River Survey
1999 - 2004
Year 0
Year 1
Year 2
Year 3
US EPA, NHEERL-WED
Corvallis, Oregon
January 31, 2000

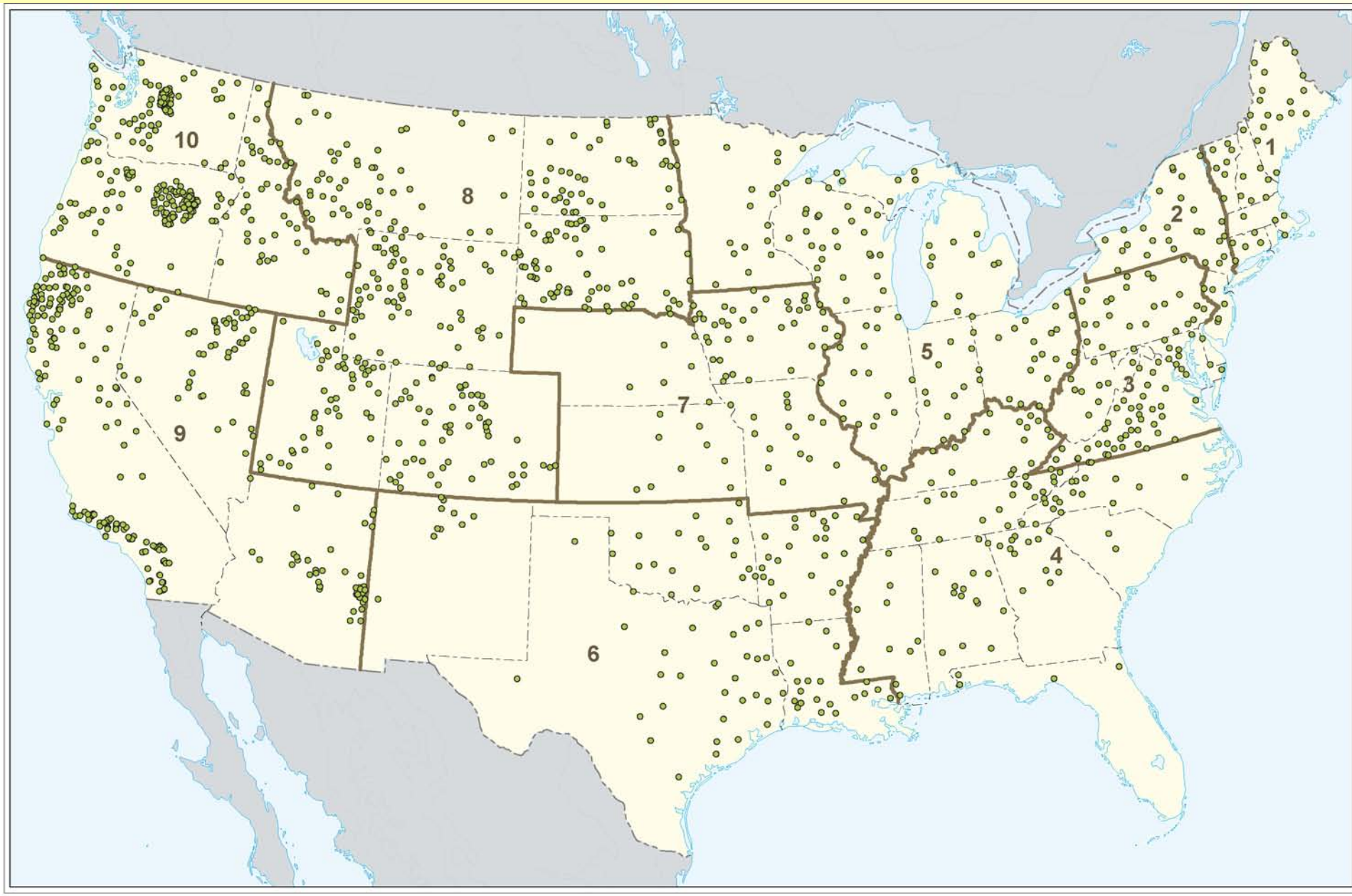
PRIMARY CANDIDATE SAMPLING SITES: 2000-2003



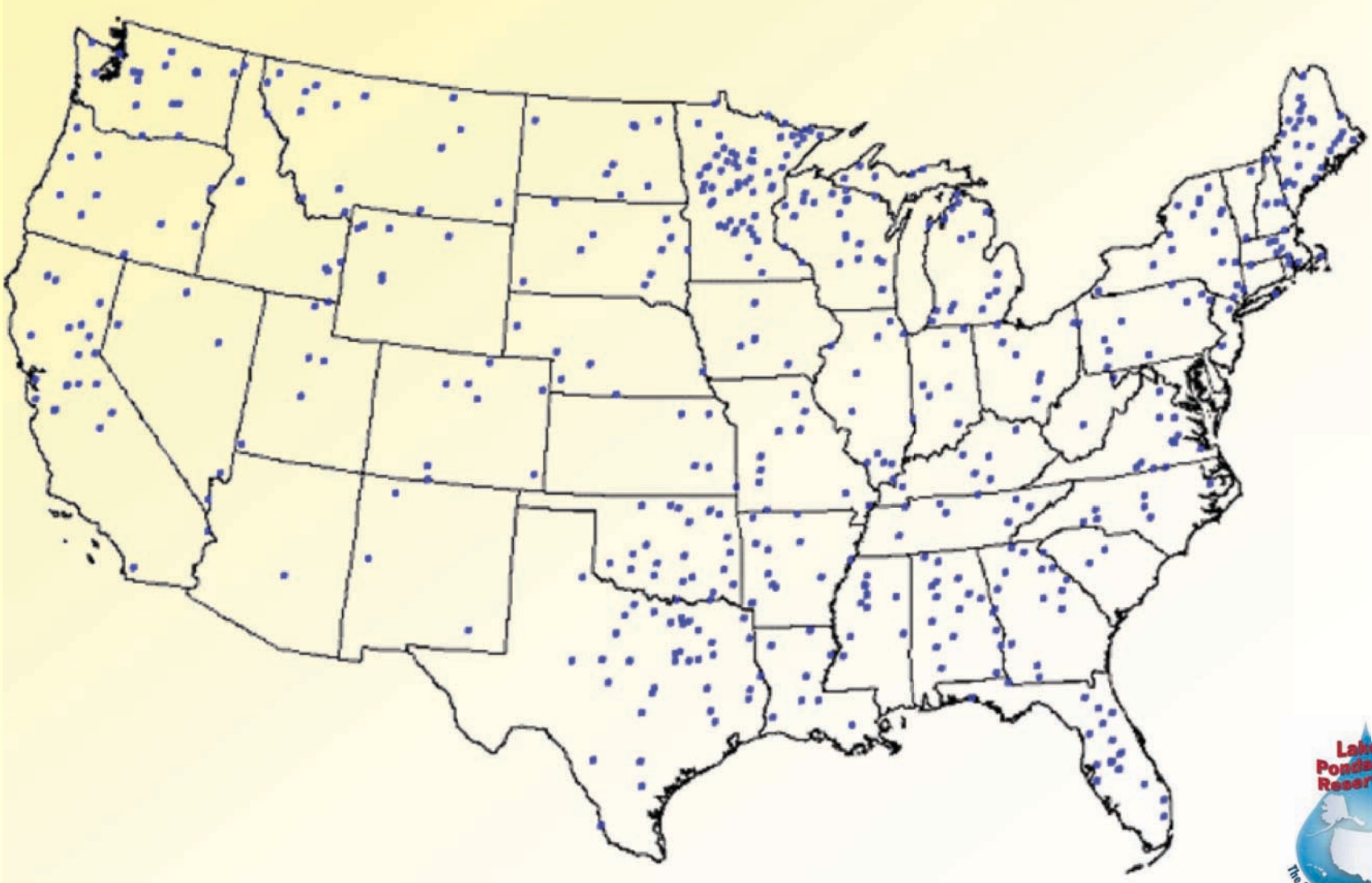
Examples -Streams



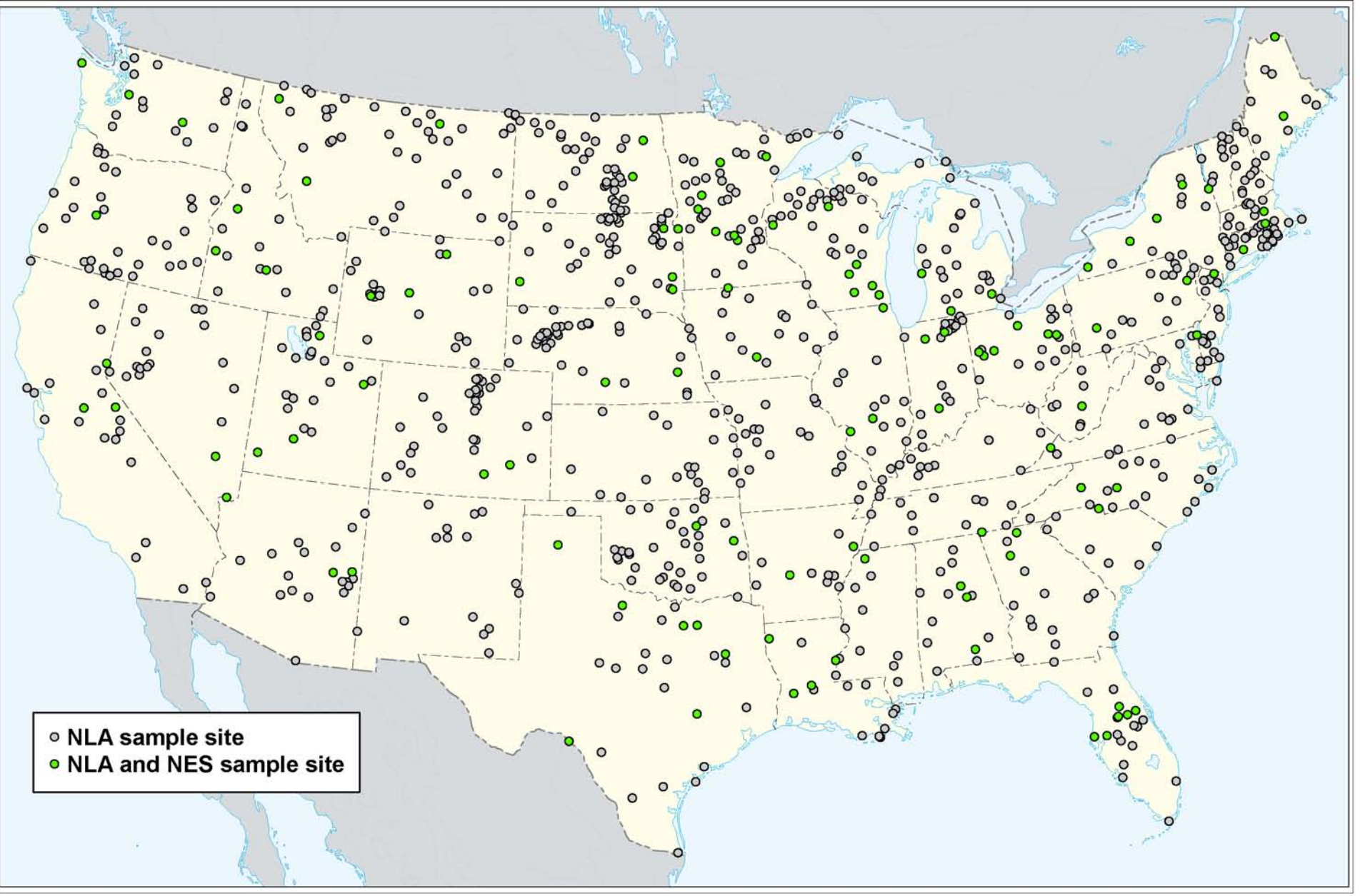
Examples - Streams



Examples - Lakes



Examples - Lakes



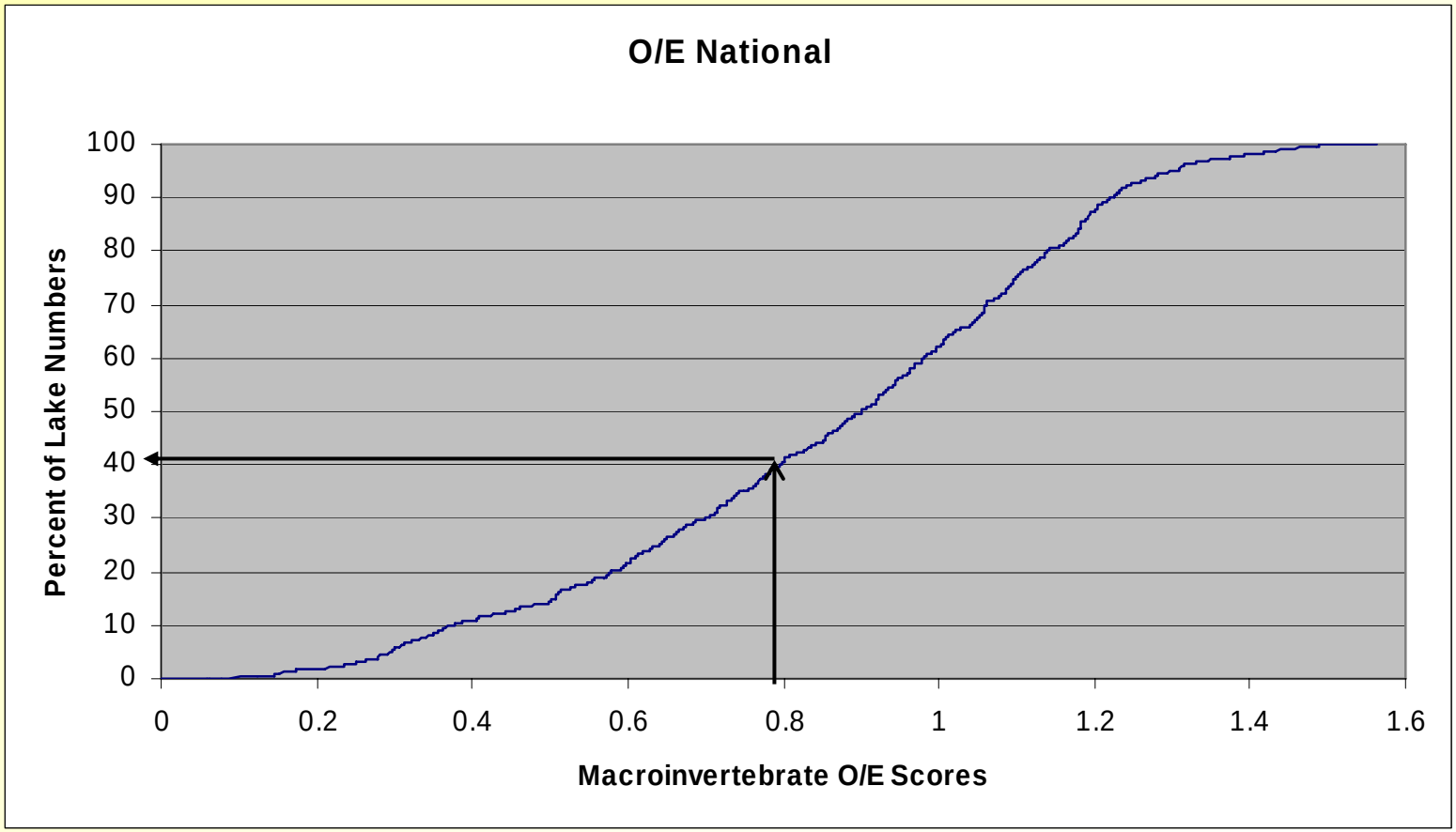
Assessment Flexibility

- Design allows post stratification as long as sample size is sufficient
- *e.g.,* Look at changes in particular parameters with various strata of interest – geographic regions, size, landcover, etc.
- Provides a “population” view of upstream-downstream questions
- Answers questions for the CWA and insights into how lakes and reservoirs function

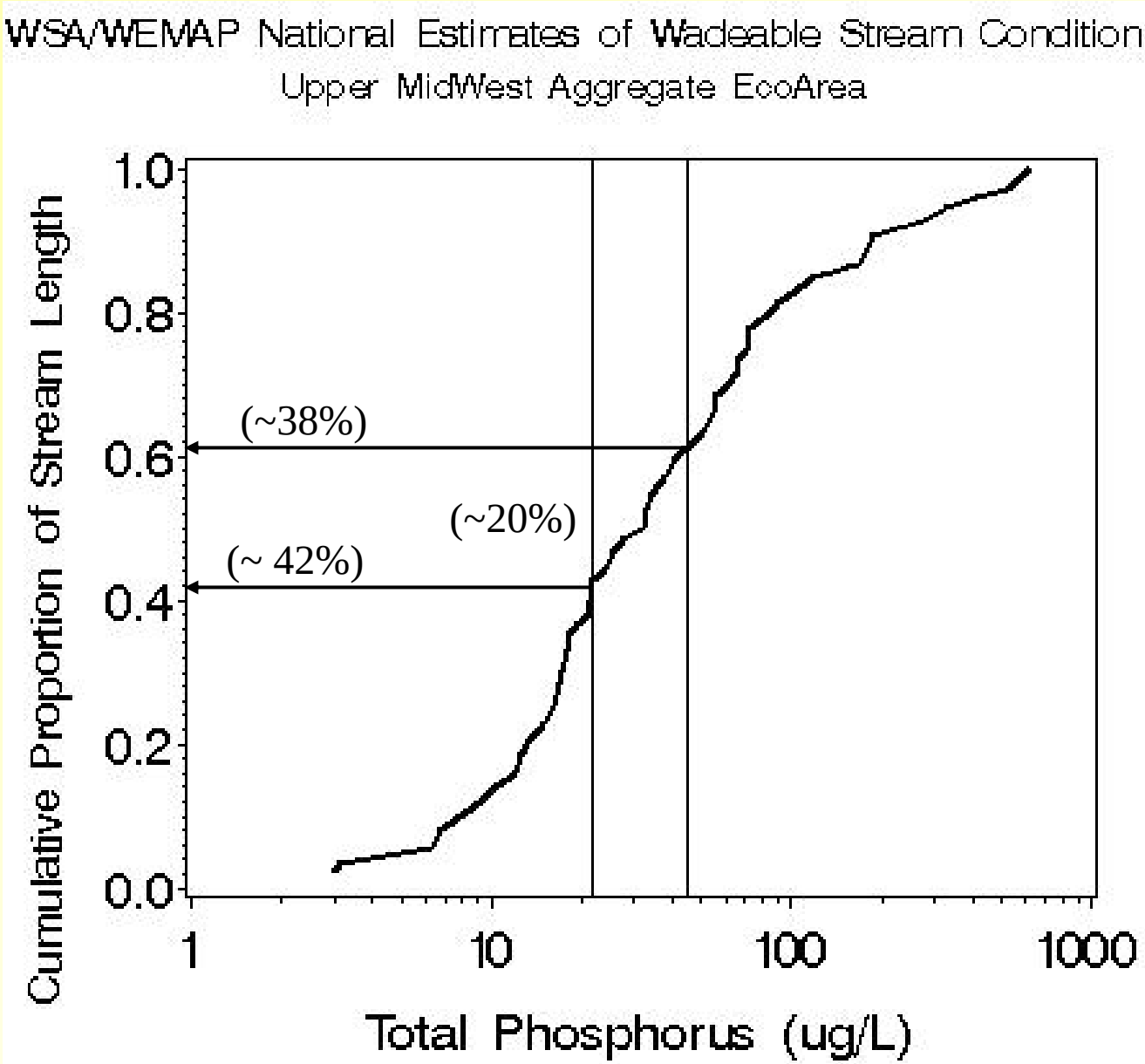
Information Derived from Survey

Distribution of Values for Indicator of Interest

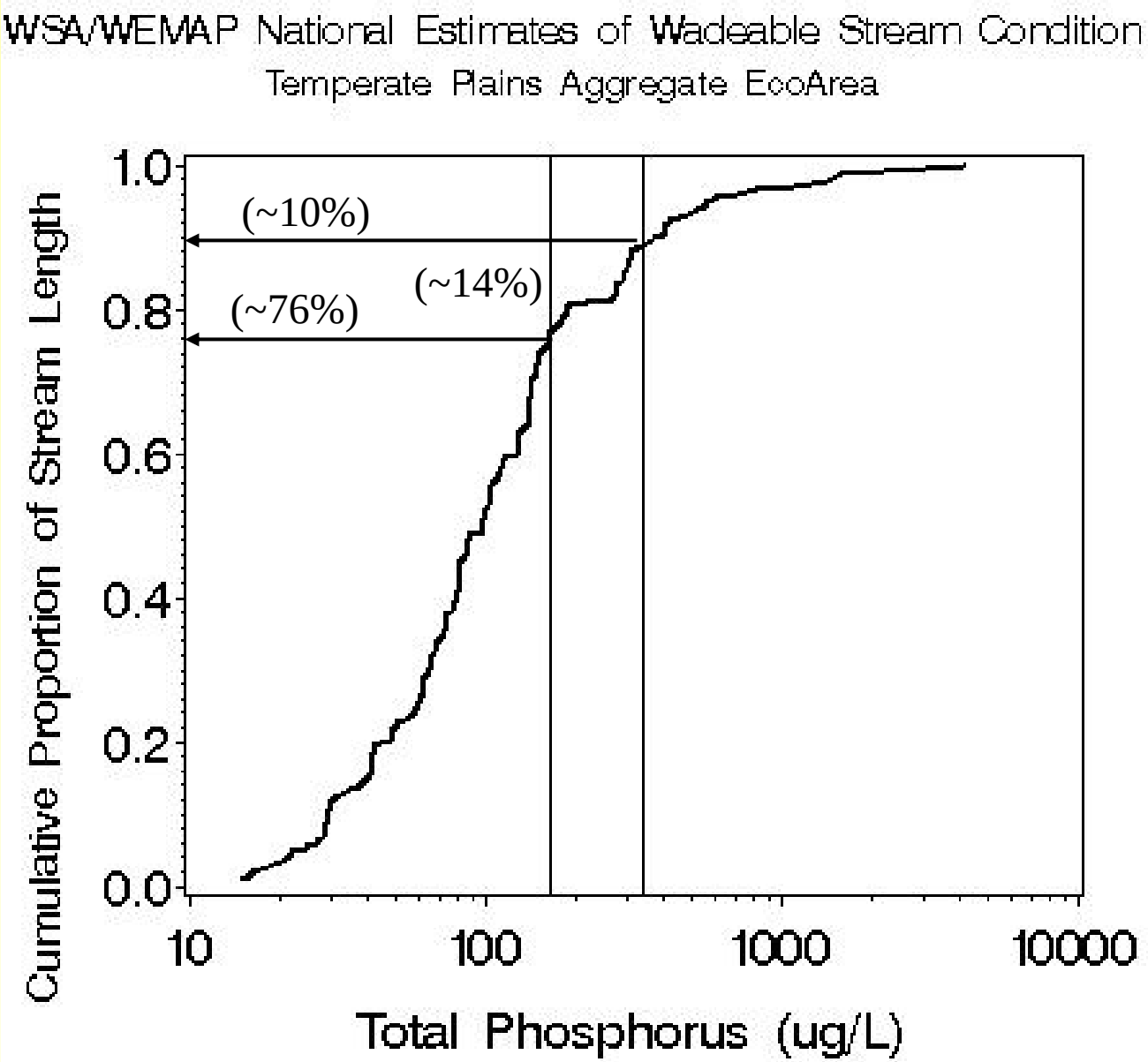
Cumulative Distribution Function (CDF)



Nutrient Distribution



Nutrient Distribution

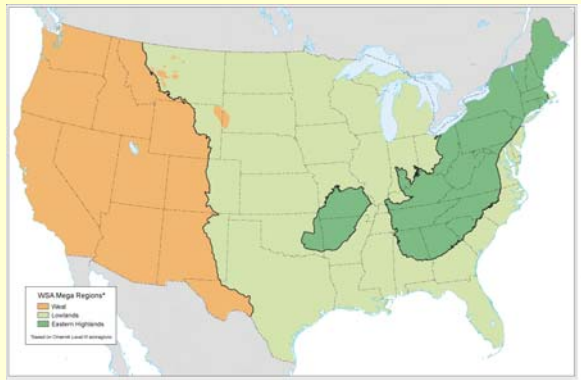


Sample Size & Uncertainty

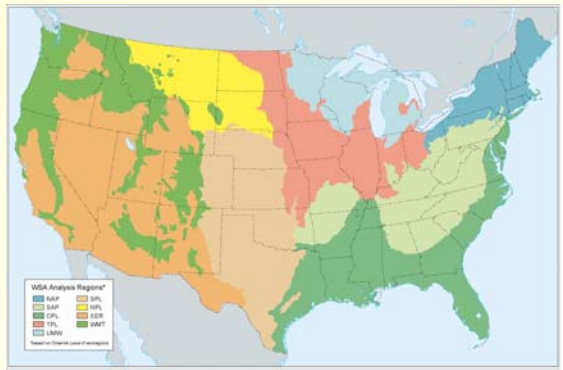
- All estimates come with confidence intervals:
  - *e.g.*, 45% of lakes have TP > 245 ug/l +/- 5%
  - Width of confidence interval driven by sample size
- Sample size of ~ 50-60 yields a confidence interval of +/- 5-8%
- Cutting the confidence interval in half requires a 4x increase in sample size.
- Reporting at multiple scales has sample size implications

Sample Size Examples

N ~ 1,300
Yield CI ~ +/- 2.5%



N ~ 400
Yield CI ~ +/- 4.6%



N ~ 75 - 150
Yield CI ~ +/- 7.5%

More Assistance

http://www.epa.gov/nheerl/arm